Modeling and Predicting Popularity Dynamics of Microblogs using Self-Excited Hawkes Processes

Peng Bao, Hua-Wei Shen, Xiaolong Jin, Xue-Qi Cheng
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
pengbaocn@gmail.com, {shenhuawei, jinxiaolong, cxq}@ict.ac.cn

ABSTRACT

The ability to model and predict the popularity dynamics of *individual* user generated items on online media has important implications in a wide range of areas. In this paper, we propose a probabilistic model using a Self-Excited Hawkes Process (SEHP) to characterize the process through which individual microblogs gain their popularity. This model explicitly captures the triggering effect of each forwarding, distinguishing itself from the reinforced Poisson process based model where all previous forwardings are simply aggregated as a single triggering effect. We validate the proposed model by applying it on Sina Weibo, the most popular microblogging network in China. Experimental results demonstrate that the SEHP model consistently outperforms the model based on reinforced Poisson process.

Categories and Subject Descriptors

J.4 [**SOCIAL AND BEHAVIORAL SCIENCES**]: Sociology; H.4 [**INFORMATION SYSTEMS APPLICATIONS**]: Miscellaneous

General Terms

Measurement; Experimentation

Keywords

popularity prediction; popularity dynamics; microblogs

1. INTRODUCTION

With the explosive growth of User Generated Contents (UGC) on online media, it becomes an important issue to predict the popularity dynamics of UGC items, including microblogs, tweets, videos, to name a few. Popularity prediction has important implications in many domains, including viral marketing, public opinion monitoring, etc. Early studies devote to characterizing the distribution of the popularity over an aggregation of UGC items [2] and making prediction by exploiting temporal correlations [1, 5].

Recently, researchers began to model the popularity dynamics of individual UGC items [3, 4]. However, these models usually assume an aggregate stochastic process without distinguishing the triggering effects of different forwarding actions in the diffusion-and-reaction process. Therefore, although these models gain success in predicting, say, the citation counts of scientific papers and view counts of Youtube videos, they are not applicable to model popularity dynamics over a microblogging network, where interactions among users matter much in popularity dynamics.

In this paper, we propose a probabilistic model using a Self-Excited Hawkes Process (SEHP) to model the process through which individual microblogs gain their popularity. This model explicitly captures the triggering effect of each forwarding, distinguishing itself from the reinforced Poisson process (RPP) based model presented in [4], where all previous forwardings are simply aggregated as a single triggering effect (see Fig. 1). We validate the proposed model by applying it on Sina Weibo[1], the most popular microblogging network in China. Experimental results demonstrate that this model consistently outperforms the model based on reinforced Poisson process.

2. THE SEHP MODEL

When a microblog spreads, it creates a *cascade* on the microblogging network. The popularity dynamics of each microblog during observed time period $[0, T]$ can be characterized by a set of time stamps t_i $(1 \leq i \leq N)$ which denote the occurrence time of each forwardings. Here, N is the total number of forwardings. Without loss of generality, we have $0 = t_0 \leq t_1 \leq t_2 \leq ... \leq t_i \leq ... \leq t_N \leq T$. For a microblog, we model its popularity dynamics using an SEHP characterized with the following rate function

$$\lambda(t) = v e^{-\beta t} + \alpha \sum_{j=1}^{j_{max}(t)} e^{-\beta(t - t_j)}, \qquad (1)$$

where v is the initial triggering strength that reflects the attractiveness of the microblog, α is the triggering strength of each subsequent forwarding, and $j_{max}(t)$ is the index of the last forwarding before time t. We set an exponential decaying function with exponent β for simplicity.

According to the survival theory, given that the $(i-1)$-th forwarding arrives at t_{i-1}, the probability that the i-th forwarding arrives at t_i follows

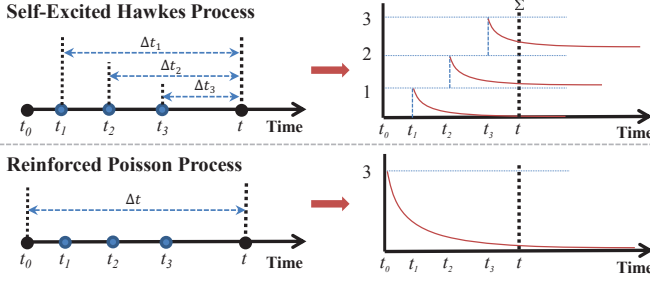$$p(t_i | t_{i-1}) = e^{-\int_{t_{i-1}}^{t_i} \lambda(t) dt} \lambda(t_i), \qquad (2)$$

[1]http://t.cn

Figure 1: Differences between SEHP and RPP

Figure 2: Prediction performance

and the probability that no forwarding arrives between t_N and T is

$$p(T|t_N) = e^{-\int_{t_N}^{T} \lambda(t)dt}. \tag{3}$$

Assuming that forwardings during different time intervals are statistically independent, the likelihood of observing a cascade of a microblog and its subsequent forwardings during time interval $[0, T]$ follows

$$\mathcal{L}(\alpha, \beta, v) = p(T|t_N) \prod_{i=1}^{N} p(t_i|t_{i-1}). \tag{4}$$

By substituting Eqs. (1), (2), and (3) in Eq. (4), we obtain the logarithmic likelihood

$$\log \mathcal{L}(\alpha, \beta, v) = \frac{v}{\beta}\left(e^{-\beta T} - 1\right) + \frac{\alpha}{\beta}\sum_{i=1}^{N}\left(e^{-\beta(T-t_i)} - 1\right) + \sum_{i=1}^{N}\log\left(ve^{-\beta t_i} + \alpha \sum_{j=1}^{j_{max}(t_i)} e^{-\beta(t_i-t_j)}\right). \tag{5}$$

We employ maximum likelihood estimation to infer the parameters in the proposed model. With the estimated parameters, the model can be used to predict the expected number $c(t)$ of forwardings of a microblog up to any given time t. With the rate function in Eq. (1), we obtain the prediction function

$$c(t) = N + \frac{v}{\beta}\left(e^{-\beta T} - e^{-\beta t}\right) + \frac{\alpha}{\beta}\sum_{j=1}^{j_{max}(t)}\left(e^{-\beta(T-t_j)} - e^{-\beta(t-t_j)}\right). \tag{6}$$

3. EXPERIMENTAL VALIDATION

Experiments are conducted on a dataset from Sina Weibo, published by the WISE 2012 Challenge[2]. We select microblogs that were submitted during July 1-31, 2011 and have more than 10 forwardings during the first hour and more than 100 forwardings during forty-eight hours after submission. This resulting dataset consists of 5670 microblogs and their cascades.

To validate the prediction performance of the SEHP, we compare it with the state-of-the-art model based on reinforced Poisson process [4], in terms of two metrics:

- Mean Absolute Percentage Error ($MAPE$): It measures the average derivation between the predicted and observed popularity over all microblogs. Denoting the predicted popularity for a microblog i up to time t as
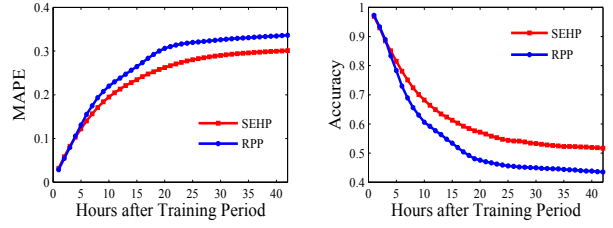
---

[2] http://www.wise2012.cs.ucy.ac.cy/challenge.html

$c_i(t)$ and its actual popularity as $r_i(t)$, the MAPE over M microblogs can be written as

$$MAPE = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{c_i(t) - r_i(t)}{r_i(t)}\right|.$$

- $Accuracy$: It measures the fraction of microblogs, correctly predicted under a given error tolerance ϵ. Specifically, the accuracy of popularity prediction over M microblogs is

$$Accuracy = \frac{1}{M}\left|\left\{i : \left|\frac{c_i(t) - r_i(t)}{r_i(t)}\right| \leq \epsilon\right\}\right|.$$

The threshold ϵ is set as 0.2 in this paper.

We set the training period, i.e., T, as 6 hours and then predict the popularity for each microblog from the 1st to 42nd hour after the training period. As shown in Fig. 2, the SEHP model consistently exhibits lower error and higher accuracy than the RPP model.

4. CONCLUSIONS

In this paper, we proposed a probabilistic model to characterize and predict the popularity dynamics of microblogs using an SEHP. Experiments on a Sina Weibo dataset demonstrated that this model consistently outperforms the baseline model based on reinforced Poisson process.

5. ACKNOWLEDGEMENTS

6. REFERENCES

[1] P. Bao, H. W. Shen, J. Huang, X. Q. Cheng. Popularity Prediction in Microblogging Network: a Case Study on Sina Weibo. In *Proc. of WWW '13*, pp. 177-178, Brazil.

[2] R. Crane, D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci.*, 105(41): 15649-15653, 2008.

[3] M. Gomez-Rodriguez, J. Leskovec, B. Schölkopf. Modeling Information Propagation with Survival Theory. In *Proc. of ICML '13*, pp. 666–674, USA.

[4] H. W. Shen, D. Wang, C. Song, A.-L. Barabási. Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes. In *Proc. of AAAI '14*, pp. 291-297, Canada.

[5] G. Szabo, B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8): 80-88, 2010.